

NAG Library Routine Document

G02CGF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of ***bold italicised*** terms and other implementation-dependent details.

1 Purpose

G02CGF performs a multiple linear regression on a set of variables whose means, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation coefficients are given.

2 Specification

```

SUBROUTINE G02CGF(N, K1, K, XBAR, SSP, LDSSP, R, LDR, RESULT, COEF,
1          LDCOE, CON, RINV, LDRINV, C, LDC, WKZ, LDWKZ, IFAIL)
    INTEGER          N, K1, K, LDSSP, LDR, LDCOE, LDRINV, LDC, LDWKZ,
1          IFAIL
    double precision XBAR(K1), SSP(LDSSP,K1), R(LDR,K1), RESULT(13),
1          COEF(LDCOE,3), CON(3), RINV(LDRINV,K), C(LDC,K),
2          WKZ(LDWKZ,K)

```

3 Description

G02CGF fits a curve of the form

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

to the data points

$$\begin{pmatrix} x_{11}, x_{21}, \dots, x_{k1}, y_1 \\ x_{12}, x_{22}, \dots, x_{k2}, y_2 \\ \vdots \\ x_{1n}, x_{2n}, \dots, x_{kn}, y_n \end{pmatrix}$$

such that

$$y_i = a + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki} + e_i, \quad i = 1, 2, \dots, n.$$

The routine calculates the regression coefficients, $b_1, b_2, \dots, b_k$, the regression constant, a , and various other statistical quantities by minimizing

$$\sum_{i=1}^n e_i^2.$$

The actual data values $(x_{1i}, x_{2i}, \dots, x_{ki}, y_i)$ are not provided as input to the routine. Instead, input consists of:

- (i) The number of cases, n , on which the regression is based.
- (ii) The total number of variables, dependent and independent, in the regression, $(k + 1)$.
- (iii) The number of independent variables in the regression, k .
- (iv) The means of all $k + 1$ variables in the regression, both the independent variables $(x_1, x_2, \dots, x_k)$ and the dependent variable (y) , which is the $(k + 1)$ th variable: i.e., $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k, \bar{y}$.

- (v) The $(k+1)$ by $(k+1)$ matrix $[S_{ij}]$ of sums of squares and cross-products of deviations from means of all the variables in the regression; the terms involving the dependent variable, y , appear in the $(k+1)$ th row and column.
- (vi) The $(k+1)$ by $(k+1)$ matrix $[R_{ij}]$ of the Pearson product-moment correlation coefficients for all the variables in the regression; the correlations involving the dependent variable, y , appear in the $(k+1)$ th row and column.

The quantities calculated are:

- (a) The inverse of the k by k partition of the matrix of correlation coefficients, $[R_{ij}]$, involving only the independent variables. The inverse is obtained using an accurate method which assumes that this sub-matrix is positive-definite.
- (b) The modified inverse matrix, $C = [c_{ij}]$, where

$$c_{ij} = \frac{R_{ij}r_{ij}}{S_{ij}}, \quad i, j = 1, 2, \dots, k,$$

where r_{ij} is the (i, j) th element of the inverse matrix of $[R_{ij}]$ as described in (a) above. Each element of C is thus the corresponding element of the matrix of correlation coefficients multiplied by the corresponding element of the inverse of this matrix, divided by the corresponding element of the matrix of sums of squares and cross-products of deviations from means.

- (c) The regression coefficients:

$$b_i = \sum_{j=1}^k c_{ij} S_{j(k+1)}, \quad i = 1, 2, \dots, k,$$

where $S_{j(k+1)}$ is the sum of cross-products of deviations from means for the independent variable x_j and the dependent variable y .

- (d) The sum of squares attributable to the regression, SSR , the sum of squares of deviations about the regression, SSD , and the total sum of squares, SST :

$SST = S_{(k+1)(k+1)}$, the sum of squares of deviations from the mean for the dependent variable, y ;

$$SSR = \sum_{j=1}^k b_j S_{j(k+1)}; \quad SSD = SST - SSR$$

- (e) The degrees of freedom attributable to the regression, DFR , the degrees of freedom of deviations about the regression, DFD , and the total degrees of freedom, DFT :

$$DFR = k; \quad DFD = n - k - 1; \quad DFT = n - 1.$$

- (f) The mean square attributable to the regression, MSR , and the mean square of deviations about the regression, MSD :

$$MSR = SSR/DFR; \quad MSD = SSD/DFD.$$

- (g) The F values for the analysis of variance:

$$F = MSR/MSD.$$

- (h) The standard error estimate:

$$s = \sqrt{MSD}.$$

- (i) The coefficient of multiple correlation, R , the coefficient of multiple determination, R^2 and the coefficient of multiple determination corrected for the degrees of freedom, $\bar{R}^2$;

$$R = \sqrt{1 - \frac{SSD}{SST}}; \quad R^2 = 1 - \frac{SSD}{SST}; \quad \bar{R}^2 = 1 - \frac{SSD \times DFT}{SST \times DFD}.$$

- (j) The standard error of the regression coefficients:

$$se(b_i) = \sqrt{MSD \times c_{ii}}, \quad i = 1, 2, \dots, k.$$

- (k) The
- t
- values for the regression coefficients:

$$t(b_i) = \frac{b_i}{se(b_i)}, \quad i = 1, 2, \dots, k.$$

- (l) The regression constant,
- a
- , its standard error,
- $se(a)$
- , and its
- t
- value,
- $t(a)$
- :

$$a = \bar{y} - \sum_{i=1}^k b_i \bar{x}_i; \quad se(a) = \sqrt{MSD \times \left(\frac{1}{n} + \sum_{i=1}^k \sum_{j=1}^k \bar{x}_i c_{ij} \bar{x}_j \right)}; \quad t(a) = \frac{a}{se(a)}.$$

4 References

Draper N R and Smith H (1985) *Applied Regression Analysis* (2nd Edition) Wiley

5 Parameters

- 1: N – INTEGER *Input*
On entry: the number of cases n , used in calculating the sums of squares and cross-products and correlation coefficients.
- 2: K1 – INTEGER *Input*
On entry: the total number of variables, independent and dependent, $(k + 1)$, in the regression.
Constraint: $2 \leq K1 < N$.
- 3: K – INTEGER *Input*
On entry: the number of independent variables k , in the regression.
Constraint: $K = K1 - 1$.
- 4: XBAR(K1) – **double precision** array *Input*
On entry: XBAR(i) must be set to $\bar{x}_i$, the mean value of the i th variable, for $i = 1, 2, \dots, k + 1$; the mean of the dependent variable must be contained in XBAR($k + 1$).
- 5: SSP(LDSSP,K1) – **double precision** array *Input*
On entry: SSP(i, j) must be set to S_{ij} , the sum of cross-products of deviations from means for the i th and j th variables, for $i, j = 1, 2, \dots, k + 1$; terms involving the dependent variable appear in row $k + 1$ and column $k + 1$.
- 6: LDSSP – INTEGER *Input*
On entry: the first dimension of the array SSP as declared in the (sub)program from which G02CGF is called.
Constraint: $LDSSP \geq K1$.
- 7: R(LDR,K1) – **double precision** array *Input*
On entry: R(i, j) must be set to R_{ij} , the Pearson product-moment correlation coefficient for the i th and j th variables, for $i, j = 1, 2, \dots, k + 1$; terms involving the dependent variable appear in row $k + 1$ and column $k + 1$.

- 8: LDR – INTEGER *Input*
On entry: the first dimension of the array R as declared in the (sub)program from which G02CGF is called.
Constraint: $LDR \geq K1$.
- 9: RESULT(13) – **double precision** array *Output*
On exit: the following information:
 RESULT(1) *SSR*, the sum of squares attributable to the regression;
 RESULT(2) *DFR*, the degrees of freedom attributable to the regression;
 RESULT(3) *MSR*, the mean square attributable to the regression;
 RESULT(4) *F*, the *F* value for the analysis of variance;
 RESULT(5) *SSD*, the sum of squares of deviations about the regression;
 RESULT(6) *DFD*, the degrees of freedom of deviations about the regression;
 RESULT(7) *MSD*, the mean square of deviations about the regression;
 RESULT(8) *SST*, the total sum of squares;
 RESULT(9) *DFT*, the total degrees of freedom;
 RESULT(10) *s*, the standard error estimate;
 RESULT(11) *R*, the coefficient of multiple correlation;
 RESULT(12) R^2 , the coefficient of multiple determination;
 RESULT(13) $\bar{R}^2$, the coefficient of multiple determination corrected for the degrees of freedom.
- 10: COEF(LDCOE,3) – **double precision** array *Output*
On exit: for $i = 1, 2, \dots, k$, the following information:
 COEF(*i*, 1)
 b_i , the regression coefficient for the *i*th variable.
 COEF(*i*, 2)
 $se(b_i)$, the standard error of the regression coefficient for the *i*th variable.
 COEF(*i*, 3)
 $t(b_i)$, the *t* value of the regression coefficient for the *i*th variable.
- 11: LDCOE – INTEGER *Input*
On entry: the first dimension of the array COEF as declared in the (sub)program from which G02CGF is called.
Constraint: $LDCOE \geq K$.
- 12: CON(3) – **double precision** array *Output*
On exit: the following information:
 CON(1) *a*, the regression constant;
 CON(2) $se(a)$, the standard error of the regression constant;
 CON(3) $t(a)$, the *t* value for the regression constant.
- 13: RINV(LDRINV,K) – **double precision** array *Output*
On exit: the inverse of the matrix of correlation coefficients for the independent variables; that is, the inverse of the matrix consisting of the first *k* rows and columns of R.
- 14: LDRINV – INTEGER *Input*
On entry: the first dimension of the array RINV as declared in the (sub)program from which G02CGF is called.
Constraint: $LDRINV \geq K$.

- 15: C(LDC,K) – **double precision** array *Output*
On exit: the modified inverse matrix, where

$$C(i, j) = R(i, j) \times RINV(i, j) / SSP(i, j), \text{ for } i, j = 1, 2, \dots, k.$$
- 16: LDC – INTEGER *Input*
On entry: the first dimension of the array C as declared in the (sub)program from which G02CGF is called.
Constraint: $LDC \geq K$.
- 17: WKZ(LDWKZ,K) – **double precision** array *Workspace*
18: LDWKZ – INTEGER *Input*
On entry: the first dimension of the array WKZ as declared in the (sub)program from which G02CGF is called.
Constraint: $LDWKZ \geq K$.
- 19: IFAIL – INTEGER *Input/Output*
On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.
On exit: IFAIL = 0 unless the routine detects an error (see Section 6).
For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, if you are not familiar with this parameter, the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

6 Error Indicators and Warnings

If on entry IFAIL = 0 or -1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Errors or warnings detected by the routine:

IFAIL = 1

On entry, $K1 < 2$.

IFAIL = 2

On entry, $K1 \neq (K + 1)$.

IFAIL = 3

On entry, $N \leq K1$.

IFAIL = 4

On entry, LDSSP < K1,
or LDR < K1,
or LDCOEF < K,
or LDRINV < K,
or LDC < K,
or LDWKZ < K.

IFAIL = 5

The k by k partition of the matrix R which is to be inverted is not positive-definite.

IFAIL = 6

The refinement following the actual inversion fails, indicating that the k by k partition of the matrix R , which is to be inverted, is ill-conditioned. The use of G02DAF, which employs a different numerical technique, may avoid this difficulty (an extra ‘variable’ representing the constant term must be introduced for G02DAF).

IFAIL = 7

Unexpected error in F04ABF.

7 Accuracy

The accuracy of any regression routine is almost entirely dependent on the accuracy of the matrix inversion method used. In G02CGF, it is the matrix of correlation coefficients rather than that of the sums of squares and cross-products of deviations from means that is inverted; this means that all terms in the matrix for inversion are of a similar order, and reduces the scope for computational error. For details on absolute accuracy, the relevant section of the document describing the inversion routine used, F04ABF, should be consulted. G02DAF uses a different method, based on F04AMF, and that routine may well prove more reliable numerically. It does not handle missing values, nor does it provide the same output as this routine. (In particular it is necessary to include explicitly the constant in the regression equation as another ‘variable’.)

If, in calculating F , $t(a)$, or any of the $t(b_i)$ (see Section 3), the numbers involved are such that the result would be outside the range of numbers which can be stored by the machine, then the answer is set to the largest quantity which can be stored as a **double precision** variable, by means of a call to X02ALF.

8 Further Comments

The time taken by G02CGF depends on k .

This routine assumes that the matrix of correlation coefficients for the independent variables in the regression is positive-definite; it fails if this is not the case.

This correlation matrix will in fact be positive-definite whenever the correlation matrix and the sums of squares and cross-products (of deviations from means) matrix have been formed either without regard to missing values, or by eliminating **completely** any cases involving missing values, for any variable. If, however, these matrices are formed by eliminating cases with missing values from only those calculations involving the variables for which the values are missing, no such statement can be made, and the correlation matrix may or may not be positive-definite. You should be aware of the possible dangers of using correlation matrices formed in this way (see the G02 Chapter Introduction), but if they nevertheless wish to carry out regression using such matrices, this routine is capable of handling the inversion of such matrices provided they are positive-definite.

If a matrix is positive-definite, its subsequent re-organisation by either G02CEF or G02CFF will not affect this property, and the new matrix can safely be used in this routine. Thus correlation matrices produced by any of G02BAF, G02BBF, G02BGF or G02BHF, even if subsequently modified by either G02CEF or G02CFF, can be handled by this routine.

It should be noted that in forming the sums of squares and cross-products matrix and the correlation matrix a column of constants should **not** be added to the data as an additional ‘variable’ in order to obtain a constant term in the regression. This routine automatically calculates the regression constant, a , and any attempt to insert such a ‘dummy variable’ is likely to cause the routine to fail.

It should also be noted that the routine requires the dependent variable to be the last of the $k + 1$ variables whose statistics are provided as input to the routine. If this variable is not correctly positioned in the original data, the means, standard deviations, sums of squares and cross-products of deviations from means, and correlation coefficients can be manipulated by using G02CEF or G02CFF to reorder the variables as necessary.

9 Example

This example reads in the means, sums of squares and cross-products of deviations from means, and correlation coefficients for three variables. A multiple linear regression is then performed with the third and final variable as the dependent variable. Finally the results are printed.

9.1 Program Text

```
*      G02CGF Example Program Text
*      Mark 14 Revised. NAG Copyright 1989.
*      .. Parameters ..
      INTEGER          K1, N, K, LDSSP, LDR, LDCOEF, LDRINV, LDC, LDWKZ
      PARAMETER        (K1=3,N=5,K=K1-1,LDSSP=K1,LDR=K1,LDCOEF=K,
+                      LDRINV=K,LDC=K,LDWKZ=K)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*      .. Local Scalars ..
      INTEGER          I, IFAIL, J
*      .. Local Arrays ..
      DOUBLE PRECISION C(LDC,K), COEF(LDCOEF,3), CON(3), R(LDR,K1),
+                      RESULT(13), RINV(LDRINV,K), SSP(LDSSP,K1),
+                      WKZ(LDWKZ,K), XBAR(K1)
*      .. External Subroutines ..
      EXTERNAL        G02CGF
*      .. Executable Statements ..
      WRITE (NOUT,*) 'G02CGF Example Program Results'
*      Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) (XBAR(I),I=1,K1), ((SSP(I,J),J=1,K1),I=1,K1),
+      ((R(I,J),J=1,K1),I=1,K1)
      WRITE (NOUT,*)
      WRITE (NOUT,*) 'Means:'
      WRITE (NOUT,99999) (I,XBAR(I),I=1,K1)
      WRITE (NOUT,*)
      WRITE (NOUT,*) 'Sums of squares and cross-products about means:'
      WRITE (NOUT,99998) (J,J=1,K1)
      WRITE (NOUT,99997) (I,(SSP(I,J),J=1,K1),I=1,K1)
      WRITE (NOUT,*)
      WRITE (NOUT,*) 'Correlation coefficients:'
      WRITE (NOUT,99998) (J,J=1,K1)
      WRITE (NOUT,99997) (I,(R(I,J),J=1,K1),I=1,K1)
      WRITE (NOUT,*)
      IFAIL = 1
*
      CALL G02CGF(N,K1,K,XBAR,SSP,LDSSP,R,LDR,RESULT,COEF,LDCOEF,CON,
+      RINV,LDRINV,C,LDC,WKZ,LDWKZ,IFAIL)
*
      IF (IFAIL.NE.0) THEN
        WRITE (NOUT,*)
        WRITE (NOUT,99996) ' ** G02CGF returned with IFAIL = ', IFAIL
      ELSE
        WRITE (NOUT,*) 'Vble      Coef      Std err      t-value'
        WRITE (NOUT,99995) (I,(COEF(I,J),J=1,3),I=1,K)
        WRITE (NOUT,*)
        WRITE (NOUT,99994) 'Const', (CON(I),I=1,3)
        WRITE (NOUT,*)
        WRITE (NOUT,*) 'Analysis of regression table :-'
        WRITE (NOUT,*)
        WRITE (NOUT,*)
        + '      Source      Sum of squares  D.F.      Mean square      F-val
+ue'
        WRITE (NOUT,*)
        WRITE (NOUT,99993) 'Due to regression', (RESULT(I),I=1,4)
        WRITE (NOUT,99993) 'About regression', (RESULT(I),I=5,7)
        WRITE (NOUT,99993) 'Total', (RESULT(I),I=8,9)
        WRITE (NOUT,*)
        WRITE (NOUT,99992) 'Standard error of estimate =', RESULT(10)
        WRITE (NOUT,99992) 'Multiple correlation (R) =', RESULT(11)
        WRITE (NOUT,99992) 'Determination (R squared) =', RESULT(12)
```

```

      WRITE (NOUT,99992) 'Corrected R squared          =', RESULT(13)
      WRITE (NOUT,*)
      WRITE (NOUT,*)
+      'Inverse of correlation matrix of independent variables:'
      WRITE (NOUT,99991) (J,J=1,K)
      WRITE (NOUT,99990) (I,(RINV(I,J),J=1,K),I=1,K)
      WRITE (NOUT,*)
      WRITE (NOUT,*) 'Modified inverse matrix:'
      WRITE (NOUT,99991) (J,J=1,K)
      WRITE (NOUT,99990) (I,(C(I,J),J=1,K),I=1,K)
      END IF
*
99999 FORMAT (1X,I4,F10.4)
99998 FORMAT (1X,3I10)
99997 FORMAT (1X,I4,3F10.4)
99996 FORMAT (1X,A,I5)
99995 FORMAT (1X,I3,3F12.4)
99994 FORMAT (1X,A,F11.4,2F13.4)
99993 FORMAT (1X,A,F14.4,F8.0,2F14.4)
99992 FORMAT (1X,A,F8.4)
99991 FORMAT (1X,2I10)
99990 FORMAT (1X,I4,2F10.4)
      END

```

9.2 Program Data

G02CGF Example Program Data

5.4000	5.8000	2.8000
99.2000	-57.6000	6.4000
-57.6000	102.8000	-29.2000
6.4000	-29.2000	14.8000
1.0000	-0.5704	0.1670
-0.5704	1.0000	-0.7486
0.1670	-0.7486	1.0000

9.3 Program Results

G02CGF Example Program Results

Means:

1	5.4000
2	5.8000
3	2.8000

Sums of squares and cross-products about means:

	1	2	3
1	99.2000	-57.6000	6.4000
2	-57.6000	102.8000	-29.2000
3	6.4000	-29.2000	14.8000

Correlation coefficients:

	1	2	3
1	1.0000	-0.5704	0.1670
2	-0.5704	1.0000	-0.7486
3	0.1670	-0.7486	1.0000

Vble	Coef	Std err	t-value
1	-0.1488	0.1937	-0.7683
2	-0.3674	0.1903	-1.9309

Const	5.7350	2.0327	2.8213
-------	--------	--------	--------

Analysis of regression table :-

Source	Sum of squares	D.F.	Mean square	F-value
Due to regression	9.7769	2.	4.8884	1.9464
About regression	5.0231	2.	2.5116	
Total	14.8000	4.		

Standard error of estimate = 1.5848
Multiple correlation (R) = 0.8128
Determination (R squared) = 0.6606
Corrected R squared = 0.3212

Inverse of correlation matrix of independent variables:

	1	2
1	1.4823	0.8455
2	0.8455	1.4823

Modified inverse matrix:

	1	2
1	0.0149	0.0084
2	0.0084	0.0144
